Doctoral Thesis Proposal

On the Privacy Implications of Real Time Bidding

Muhammad Ahmad Bashir
College of Computer and Information Science
Northeastern University
ahmad@ccs.neu.edu

November 25, 2018

Abstract

The massive growth of online advertising has created a need for commensurate amounts of user tracking. Advertising companies track online users extensively to serve them targeted advertisements. On the surface, this seems like a simple process: a tracker places a unique cookie in the user's browser, repeatedly observes the same cookie as the user surfs the web, and finally uses the accrued data to select targeted ads.

However, the reality is much more complex. The rise of Real Time Bidding (RTB) has forced advertising companies to collaborate more closely with each other via *cookie matching*. Because of RTB, tracking data is not just observed by trackers embedded directly into web pages, but rather it is funneled through the advertising ecosystem through complex networks of exchanges and auctions. Additionally, to gain a complete picture of user's browsing behavior and interests across all devices (e.g., laptops, smartphones, IoT devices, *etc.*), Advertising and Analytics (*A&A*) companies actively try to link all devices associated with a user through cross-device tracking.

Numerous surveys have shown that web users are not completely aware of the amount of data sharing that occurs between A&A companies, and thus underestimate the privacy risks associated with online tracking. In order to quantify users' true *digital footprints*, we need to take into account information sharing during RTB and cross-device tracking. However, measuring these flows of tracking information is challenging. Although there has been recent work on detecting information sharing (*cookie matching*) between ad exchanges, these studies are based on brittle heuristics that cannot detect all forms of information sharing, especially under adversarial conditions (e.g., obfuscation). Furthermore, since tracking mechanisms vary across different devices, these studies cannot be effectively used to study cross-device tracking. This limits our view of the privacy landscape and hinders the development of effective privacy tools.

In this thesis, I propose a content-agnostic methodology that is able to detect client- and server-side information flows between arbitrary ad exchanges using *retargeted ads*. Intuitively, this methodology works because it relies on the *semantics* of how exchanges serve ads, rather than focusing on specific cookie matching *mechanisms*. Using crawled data on 35,448 ad impressions, we show that this methodology can successfully categorize four different kinds of information sharing behavior between ad exchanges, including cases where existing heuristic methods fail.

Since our methodology does not look for patterns or identifiers in network traffic, but rather relies on causal inference, I plan to use it to understand cross-device tracking. By conducting controlled experiments, I propose to investigate which ad exchanges are involved in cross-device tracking and which identifiers they leverage to track users across devices.

Our methods allow us to collect a novel and accurate dataset of the relationships between online advertisers and trackers. Using this dataset, I propose to investigate the privacy implications of ubiquitous online tracking. Our data can be used to represent the online advertising ecosystem as a graph; on this graph I plan to run simulations to understand the diffusion of users' tracking data across the advertising ecosystem. These simulations will allow us to quantify users' true *digital footprints*, as well as evaluate the relative effectiveness of privacy preserving tools (e.g., ad and tracker blockers).

The overall goal of my thesis is to bridge the divide between the *actual* privacy landscape and our understanding of it. My thesis proposes techniques that will help

provide users with a more realistic view of the online advertising ecosystem, and enable them to gain a more accurate view of their *digital footprint*. Furthermore, the results from this thesis can be used to build better or enhance existing privacy preserving tools.

Contents

1 Introduction **1**

2 Background and Definitions **4**
2.1 Online Display Advertising 4
2.2 Targeted Advertising 5

3 Related Work **7**
3.1 The Online Advertising Ecosystem 7
3.2 Online Tracking 7
3.3 Real Time Bidding and Cookie Matching 8

4 Detecting Information Sharing Between Ad Exchanges **9**
4.1 Approach 9
4.2 Creating Shopper Personas 11
4.3 Ad Collection 11
4.4 Identifying Retargeted Ads 12
4.5 Information Flow Categorization 14
4.5.1 Categorization Rules 15
4.6 Results 16

5 Cross-Device Tracking **18**

6 Information Diffusion in the Advertising Graph **19**

7 Towards Automation of Privacy Protection Filter Lists **21**

8 Research Plan **22**

1 Introduction

The online display advertising industry has seen massive growth in the last decade. In 2017, $83B was spent on digital advertising in the U.S., and double-digit growth is forecast for each subsequent year [4]. This increased spending is fueled by advances in the industry that enable advertisers to track and target users with increasing levels of precision.

People have complicated feelings with respect to online behavioral advertising. While surveys have shown that some users prefer relevant, targeted ads to random, untargeted ads [86, 22], this preference has caveats. For example, users are uncomfortable with ads that are targeted based on sensitive Personally Identifiable Information (PII) [59, 9] or specific kinds of browsing history (e.g., visiting medical websites) [55]. Furthermore, some users are universally opposed to online tracking, regardless of circumstance [62, 86, 22].

One particular concern held by users is their *digital footprint* [43, 90, 83], i.e., which first- and third-parties are able to track their browsing history? Large-scale web crawls have repeatedly shown that trackers are ubiquitous [37, 33], with DoubleClick alone being able to observe visitors on 40% of websites in the Alexa Top-100K [18]. These results paint a picture of a balkanized web, where trackers divide up the space and compete for the ability to collect data and serve targeted ads.

However, this picture of the privacy landscape is at odds with the current reality of the ad ecosystem. Recently, the online display advertising ecosystem has seen a shift from *ad networks* to *ad exchanges*, where advertisers bid on *impressions* being sold in Real Time Bidding (RTB) auctions. RTB currently holds a 30% share of digital advertising spending in the U.S. [2] and this share is forecast to be 80% by 2022 [1]. The rise of RTB has forced advertisers to collaborate more closely with one another by sharing unique user identifiers through *cookie matching*, which is a pre-condition for advertisers to participate in RTB auctions. Due to this close collaboration among advertisers and ad exchanges, we cannot view them as isolated islands of data anymore. To capture a more realistic picture of the privacy landscape, we have to take into account the information sharing between all Advertising and Analytics (A&A) companies.

Another major change in the advertising ecosystem has occurred due to the change in internet access pattern. Users no longer rely on just desktop devices to access the internet, but rather employ multiple devices to browse the web and use application. Almost 77% of Americans now own a smartphone [21], and more than 40% use multiple devices to access the internet [30]. This change in internet access patterns has divided user activity over multiple devices and pushed advertisers to track users across all of these devices. To maximize the information about a particular user and gain a complete picture of their interests, advertisers attempt to correlate the tracking data from all devices used by a particular user through techniques broadly known as *cross-device tracking*.

Cross-device tracking has made the advertising ecosystem even more complex. Besides two studies [16, 92] which highlight the potential and existence of cross-device tracking, we currently do not understand *which* A&A companies share information across devices and *what* mechanisms they use to link the devices for a particular user. Through RTB, user tracking data is further shared with advertising partners. While some users are aware that they are being tracked online [86, 22], they might not necessarily know how much and how often their information share hands due to RTB and *cookie matching*. Due to this lack of

information, we under-estimate the privacy *digital footprint* of the user, which, in turn affects the development of effective privacy tools.

By understanding how the modern advertising ecosystem works, while taking into account the effects of RTB and cross-device tracking, we may be able to develop better privacy tools for users. These tools could bring more transparency to the complex advertising industry and give more control to users over their privacy. For example, as shown in [59, 9], some users are uncomfortable with ads that are targeted based on sensitive Personally Identifiable Information (PII). With better privacy tools, these users would be able to control the attributes advertisers use to target them.

Despite the pressing need to understand the complexities of the advertising ecosystem, we currently lack the tools to fully understand how information is being shared between A&A companies. Although there has been prior empirical work on detecting information sharing between ad exchanges [5, 72, 35], these works rely on heuristics that look for specific strings in HTTP messages to identify cookie matching. These heuristics are brittle in the face of obfuscation: for example, DoubleClick cryptographically hashes their cookies before sending them to other advertising partners [3]. More fundamentally, analysis of *client-side* HTTP messages is insufficient to detect *server-side* information flows between ad exchanges: for example, two advertisers belonging to the same parent company can share user tracking data without cookie matching. We demonstrate in § 4.6 that heuristics from prior work can miss up to 31% of cookie matches. Similarly, these heuristics fail to identify information sharing on the server-side among Google services.

Since tracking mechanisms vary across different devices, detecting information sharing between A&A companies becomes even more challenging when it comes to cross-device tracking. For example, cookies are used to track users on desktop devices, while advertising IDs are used to track users on mobile apps. Because of these different tracking mechanisms, we cannot rely on detecting specific patterns in HTTP traffic, as proposed by prior work [5, 72, 35], to study cross-device tracking.

In my thesis, I propose to bridge the divide between the *actual* privacy landscape and our understanding of it. The goal of my work is to come up with techniques and tools that will help provide a more realistic view of the online advertising ecosystem, and enable users gain a more accurate view of their *digital footprint*.

Detecting Information Flows. To this end, in § 4, I propose a novel methodology that is able to detect client- and server-side flows of information between arbitrary ad exchanges that serve *retargeted ads*. Retargeted ads are the most specific form of behavioral ads, where a user is targeted with ads related to the exact products she has previously browsed (see § 2.2 for definition). For example, Bob visits `nike.com` and browses for running shoes but decides not to purchase them. Bob later visits `cnn.com` and sees an ad for the exact same running shoes from Nike.

Our key insight is to leverage retargeted ads as a mechanism for identifying information flows between arbitrary ad exchanges. This is possible because the strict conditions that must be met for a retargeted ad to be served allow us to infer the precise flow of tracking information that facilitated the serving of the ad. Intuitively, our methodology works because it relies on the *semantics* of how exchanges serve ads, rather than focusing on specific cookie matching *mechanisms*. Specifically, instead of relying on HTTP messages to detect cookie

matching, we rely on causality; i.e., if ad network a_1 observes user u browsing a product p on shop s and if later a_1 serves u a retargeted ad for p after winning the RTB auction held by ad exchange e_1, then it implies that e_1 and a_1 have shared user identifiers. Otherwise, a_1 would have no way of identifying u as the source of the impression (see § 2.2) during RTB and would not pay the premium price to win the auction.

We demonstrate the efficacy of our methodology by conducting extensive experiments on real data. We train 90 *personas* by visiting popular e-commerce sites (§ 4.2), and then crawl major publishers to gather retargeted ads [13, 19]. To record detailed information about the provenance of third-party resource inclusions in webpages (i.e., which resource included which other resource), all crawls were performed using an instrumented version of Chromium [11] that records the *inclusion chain* for every resource it encounters, including 35,448 chains associated with 5,102 unique retargeted ads (§ 4.1).

We use carefully designed pattern matching rules in § 4.5 to categorize each of the retargeted ad chains into four different categories, which reveal 1) the pair of ad exchanges that shared information in order to serve the retarget, and 2) the mechanism they used to share the data (e.g., cookie matching).

Since our methodology is platform-agnostic, we provided empirical evidence that Google shares data across its services by detecting server-side information flows. Furthermore, in total, our methodology identified 200 cookie matching pairs, out of which 31% were missed by heuristic methods used by prior works to analyze cookie matching.

This work has been completed and was published at *USENIX Security* in 2016.

Cross-Device Tracking. Now that we have a methodology that can detect information flows between arbitrary ad exchanges that is mechanism agnostic, I propose to use it to study cross-device tracking. The high-level methodology remains the same: train personas and elicit retargeted ads to causally identify flows of tracking data. The only difference is that instead of creating personas and collecting ads within a single desktop web browser, we create personas and browse products on one physical device (e.g., a desktop), and then visit publishers to collect ads on another physical device (e.g., a smartphone).

I plan on analyzing the inclusion chains produced during our experiments to detect *which* ad exchanges share tracking data across devices. By conducting controlled experiments and leaking *only one* identifier to the ad networks during the product browsing stage, we can identify *which* identifier was used for cross-device tracking. I plan on investigating the usage of both deterministic (e.g., email address, advertsing ID, IMEI) and probabilistic (e.g., location, Wifi network, browsing behavior) identifiers for cross-device tracking.

This project poses several challanges. First of all, we need to instrument all Android devices to control all the identifiers. In addition to that, mobile Chrome is more complicated to debug and to extract inclusion chains from. Furthermore, we need to carefully construct personas on mobile devices and simulate the browsing behavior through several well-tested apps.

I have completed the initial test harness to conduct experiments for this project. I have also finalized the experiments to check the usage of IP address for the purpose of cross-device tracking. I aim to wrap up this project in early 2019.

Information Diffusion. To demonstrate the effect of Real Time Bidding (RTB) on users' *digital footprint* (i.e., the portion of their browsing that is observable to A&A compa-

nies), I propose to simulate users' browsing behavior and their tracking information diffusion in the advertising ecosystem. To this end, I use the *accurate* information flows between ad exchanges I collected to model the advertising ecosystem in the form of a graph called an *Inclusion* graph. By simulating browsing traces for 200 users based on empirical data, we show that the *Inclusion* graph can be used to model the diffusion of user tracking data across the advertising ecosystem.

Our results demonstrate that due to RTB, the major A&A companies observe the vast majority of users' browsing history. Even under realistic conditions where only a small number (37) of well-connected ad exchanges indirectly share impressions during RTB auctions, the top 10% of A&A companies observe more than 91% impressions (page visits) and 82% visited publishers.

Furthermore, by simulating the effect of five blocking strategies, we find that AdBlock Plus (the world's most popular ad blocking browser extension [75, 60], is ineffective at protecting users' privacy because major ad exchanges are whitelisted under the Acceptable Ads program [88]. In contrast, Disconnect blocks the most information flows to advertising companies, followed by removal of top 10% A&A companies. However, even with strong blocking, major ad exchanges still observe 40–70% of user impressions.

This project is currently under submission at *The Privacy Enhancing Technologies Symposium* (PETS).

Outline. The rest of my thesis proposal is structured as follows: I begin by providing background and introducing definitions in § 2. Then I provide a comprehensive survey of related work in § 3. Next, I describe in § 4 the methodology I developed to study information flows using retargeted ads. In § 5, I propose how to use this methodology to detect cross-device tracking. Using the information flows I have collected, I present in § 6 how user tracking information gets diffused in the advertising ecosystem. I conclude by describing my research plan in § 8.

2 Background and Definitions

In this section, I provide essential background information and definitions about the online advertising industry that will be used throughout this proposal.

2.1 Online Display Advertising

Online display advertising is fundamentally a matching problem. On one side are *publishers* (e.g., news websites, blogs, *etc.*) who produce content, and earn revenue by displaying ads to users. On the other side are advertisers who want to display ads to particular users (e.g., based on demographics or market segments). Unfortunately, the online user population is fragmented across hundreds of thousands of publishers, making it difficult for advertisers to reach desired customers.

Ad networks bridge this gap by aggregating *inventory* from publishers (i.e., space for displaying ads) and filling it with ads from advertisers. Ad networks make it possible for advertisers to reach a broad swath of users, while also guaranteeing a steady stream of revenue for publishers. Inventory is typically sold using a Cost per Mille (CPM) model,

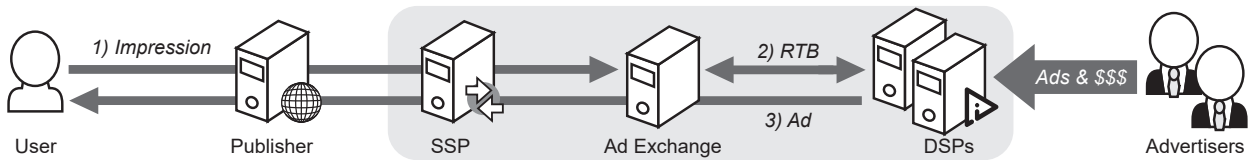

Figure 1: The display advertising ecosystem. Impressions and tracking data flow left-to-right, while revenue and ads flow right-to-left.

where advertisers purchase blocks of 1000 *impressions* (views of ads), or a Cost per Click (CPC) model, where the advertiser pays a small fee each time their ad is clicked by a user. I collectively refer to companies engaged in analytics and advertising as *A&A* companies.

Ad Exchanges and Auctions. Over time, ad networks are being supplanted by *ad exchanges* that rely on an auction-based model. In Real-time Bidding (*RTB*) exchanges, advertisers bid on individual impressions in real-time; the winner of the auction is permitted to serve an ad to the user. Google's DoubleClick is the largest ad exchange, and it supports RTB [3].

As shown in Figure 1, there is a distinction between Supply-side Platforms (*SSPs*) and Demand-side Platforms (*DSPs*) with respect to ad auctions. SSPs work closely with publishers to manage their relationships with multiple ad exchanges, typically to maximize revenue by forwarding impression inventory to the most lucrative ad exchange. For example, OpenX is an SSP [73]. In contrast, DSPs work with advertisers to assess the value of each impression and optimize bid prices. MediaMath is an example of a DSP [63]. To make matters more complicated, many companies offer products that cross categories; for example, Rubicon Project offers SSP, ad exchange, and DSP products [81]. A more detailed discussion of the modern online advertising ecosystem can be found in [61].

2.2 Targeted Advertising

Initially, the online display ad industry focused on generic brand ads (e.g., "Enjoy Coca-Cola!") or *contextual ads* (e.g., an ad for Microsoft on StackOverflow). However, the industry quickly evolved towards *behaviorally targeted ads* that are served to specific users based on their browsing history, interests, and demographics.

Tracking. To serve targeted ads, ad exchanges and advertisers must collect data about online users by tracking their actions. Publishers embed JavaScript or invisible "tracking pixels" that are hosted by tracking companies into their web pages, thus any user who visits the publisher also receives third-party cookies from the tracker (other, more sophisticated tracking mechanisms are discussed in § 3). Numerous studies have shown that trackers are pervasive across the web [52, 50, 80, 18, 32, 45], which allows advertisers to collect users' browsing history. All major ad exchanges, like DoubleClick and Rubicon, perform user tracking, but there are also companies like BlueKai that just specialize in tracking.

Cross-device Tracking. Recently, there has been a major shift by users towards mobile device usage. According to a recent study, 77% of Americans now own a smartphone [21], and more than 40% use multiple devices to access the internet [30]. To gain a complete picture of a user's interests, A&A companies have started tracking users across multiple devices.

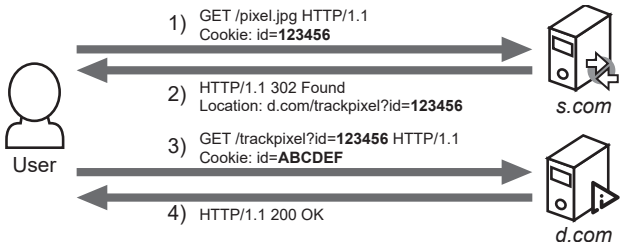

Figure 2: SSP s matches their cookie to DSP d using an HTTP redirect. Notice how *s.com*'s cookie is sent to *d.com* as request parameter.

This technique, known as cross-device tracking, consists of two approaches: deterministic and probabilistic. Deterministic cross-device tracking relies on identifying a user based on some PII like email address that is present across all of a user's devices. For example, Google and Facebook are able to track users' across devices because users typically log-in to both of these platforms on each device they use. Probabilistic cross-device tracking, on the other hand, attempts to identify a user based on several non-specific data points like Wifi network SSIDs, geolocation, IP address, browsing patterns, *etc.*. By clustering these data points from different devices, clusters may emerge that tie the devices, and thus their owner, together.

Cookie Matching. During an RTB ad auction, DSPs submit bids on an impression. The amount that a DSP bids on a given impression is intrinsically linked to the amount of information they have about that user. For example, a DSP is unlikely to bid highly for user u whom they have never observed before, whereas a DSP may bid heavily for user v who they have recently observed browsing high-value websites (e.g., the baby site `TheBump.com`).

However, the Same Origin Policy (SOP) [66] hinders the ability of DSPs to identify users in ad auctions. As shown in Figure 1, requests are first sent to an SSP which forwards the impression to an exchange (or holds the auctions itself). At this point, the SSP/exchange's cookies are known, but not the DSPs. This leads to a catch-22 situation: a DSP cannot read its cookies until it contacts the user, but it cannot contact the user without first bidding and winning the auction.

To circumvent SOP restrictions, ad exchanges and advertisers engage in *cookie matching* (sometimes called *cookie syncing*). Cookie matching is illustrated in Figure 2: the user's browser first contacts ad exchange `s.com`, which returns an HTTP redirect to its partner `d.com`. `s` reads its own cookie, and includes it as a parameter in the redirect to `d`. `d` now has a mapping from its cookie to `s`'s. In the future, if `d` participates in an auction held by `s`, it will be able to identify matched users using `s`'s cookie. Note that some ad exchanges (including DoubleClick) send cryptographically hashed cookies to their partners, which prevents the ad network's true cookies from leaking to third-parties.

Retargeted Ads. In this thesis, I propose a methodology that leverages *retargeted ads*, which are the most specific type of targeted display ads. Two conditions must be met for a DSP to serve a retargeted ad to a user u: 1) the DSP must know that u browsed a specific product on a specific e-commerce site, and 2) the DSP must be able to uniquely identify u during an auction. If these conditions are met, the DSP can serve u a highly personalized ad reminding them to purchase the product from the retailer. Cookie matching is crucial for ad retargeting, since it enables DSPs to meet requirement (2).

3 Related Work

In this section, I survey related work on the structure of the online advertising ecosystem, mechanisms for online tracking, and work that has specifically examined RTB and cookie matching.

3.1 The Online Advertising Ecosystem

Numerous studies have chronicled the online advertising ecosystem, which is composed of companies that: track users, serve ads, act as platforms between *publishers* (websites that rely on advertising revenue to pay for content creation) and advertisers, or all of the above. Mayer et al. present an accessible introduction to this topic in [61]. Barford et al. [13] looked at the major ad networks, targeted ads, and associated user characteristics on the web by mapping the online *adscape*, whereas Rodriguez et al. measured the ad ecosystem on mobile devices [87]. More recently, Razaghpanah et al. [76] presented insights into the mobile advertising and tracking ecosystem. Gill et al. [37] used browsing traces to study the economy of online advertising and discovered that the revenues are skewed towards the largest trackers (primarily Google). Acar et al. [5] conducted crawls over Alexa Top-3K to find user identifiers being shared across domains. More recently, Cahn et al. performed a broad survey of cookie characteristics across the web, and found that <1% of trackers can aggregate information across 75% of websites in the Alexa Top-10K [18]. Falahrastegar et al. [34] expanded on these results by comparing trackers across geographic regions, while Li et al. showed that most tracking cookies can be automatically detected using simple machine learning methods [57].

Other empirical studies have focused more on the individual implications of targeted advertising. Guha et al. [41] developed a controlled and systematic method for measuring online ads on the web based on trained *personas*. Carrascosa et al. [19] used these methods to prove that advertisers use sensitive attributes about users when targeting ads.

Researchers have also studied malicious and bad practices in the advertising ecosystem. Zarras et al. [91] studied malicious ad campaigns and the ad networks associated with them, whereas Bashir et al. [14] found that some advertisers were not following industry guidelines and were serving poor quality ads.

3.2 Online Tracking

Tracking mechanisms. To facilitate ad targeting, participants in the ad ecosystem must extensively track users. Krishnamurthy et al. were one of the first to bring attention to the pervasiveness of trackers and their privacy implications for users [52], and since then they have been cataloging the spread of trackers and assessing the ensuing privacy implications [49, 50, 51]. Recently, Lerner et al. [56] examined the evolution of tracking over time.

Advertisers have changed their tracking techniques over time, sometimes going to extraordinary lengths to collect and retain user information. Some of the techniques they employ involve leveraging persistent cookies [47], local state in browser plugins [82, 12], browsing history through extensions [84], and browser fingerprinting methods [65, 71, 80, 69, 5, 48, 33].

Recently, Englehardt et al. [32] found trackers fingerprinting users via the JavaScript `Audio` and `Battery Status` APIs.

Researchers have also studied the state of tracking and its privacy implications on mobile devices [87, 15, 31, 40, 78, 76]. They have noticed that tracking is ubiquitous on mobile devices and that apps use embedded sensors (e.g., camera, microphone, GPS) to extensively track users.

Additionally, there have been two prominent studies on cross-device tracking. Brookman et al. [16] from the Federal Trade Commission (FTC) surveyed 100 popular websites to study the potential for cross-device tracking, although they did not measure the actual prevalence of cross-device tracking. In contrast, Zimmeck et al. [92] found empirical evidence of cross-device tracking in their survey of 126 internet users.

User Profiles. Several studies specifically focus on tracking data collected by Google, since their trackers are more pervasive than any others on the web [37, 18]. Alarmingly, two studies have found that Google's Ad Preferences Manager, which is supposed to allow users to see and adjust how they are being targeted for ads, actually hides sensitive information from users [89, 24]. This finding is troubling given that several studies rely on data from the Ad Preferences Manager as their source of ground-truth [41, 20, 13]. To combat this lack of transparency, Lecuyer et al. [53, 54] have built systems that rely on controlled experiments and statistical analysis to infer the profiles that Google constructs about users. Castelluccia et al. [20] go further by showing that adversaries can infer users' profiles by passively observing the targeted ads they are shown by Google.

User Concerns and Reaction. On one hand, tracking has enabled advertisers to show relevant ads to users, while on the other, it has raised concerns among users about the amounts and types of information being collected about them [62, 85]. To avoid pervasive tracking, users are increasingly adopting tools that block trackers and ads [75, 60]. There has also been development towards whitelisting "acceptable" ads [88].

Concerned with the increased adoption of ad and tracker blocking tools, advertisers have started developing techniques to counter them. Merzdovnik et al. critically examined the effectiveness of tracker blocking tools [64]; in contrast, Nithyanand et al. studied advertisers' efforts to counter ad blockers [70]. Mughees et al. examined the prevalence of anti-ad blockers in the wild [67]. Recently, advertisers were reported by user communities for displaying ads (even with ad blockers installed) through WebSockets and WebRTC [42, 79]. Similarly, WebRTC has also been known to reveal user IP addresses [32, 77].

The research community has proposed a variety of mechanisms to stop online tracking that go beyond blacklists of domains and URLs. Li et al. [57] and Ikram et al. [44] used machine learning to identify trackers, while Papaodyssefs et al. [74] examined the use of private cookies to avoid being tracked. Nikiforakis et al. propose the complementary idea of adding entropy to the browser to evade fingerprinting [68]. However, despite these efforts, third-party trackers are still pervasive and pose real privacy issues to users [64].

3.3 Real Time Bidding and Cookie Matching

As we note in § 2, advertisers have to perform *cookie matching* to be able to participate in RTB auctions. Although ad networks have been transitioning to RTB auctions since the mid-

2000s, only three empirical studies have examined cookie matching. Acar et al. found that hundreds of domains passed unique identifiers to each other while crawling websites in the Alexa Top-3K [5]. Olejnik et al. noticed that ad auctions were leaking the winning bid prices for impressions, thus enabling a fascinating behind-the-scenes look at RTB auctions [72]. In addition to examining the monetary aspects of auctions, Olejnik et al. found 125 ad exchanges using cookie matching. Finally, Falahrastegar et al. examine the clusters of domains that all share unique, matched cookies using crowdsourced browsing data [35].

Retargeted Ads. Several studies have examined retargeted ads, which are directly facilitated by cookie matching and RTB. Liu et al. identified and measured retargeted ads served by DoubleClick by relying on unique AdSense tags that were embedded in ad URLs [58]. Olejnik et al. crawled specific e-commerce sites to elicit retargeted ads from those retailers, and observed that retargeted ads could cost advertisers over $1 per impression (an enormous sum, considering contextual ads sell for <$0.01) [72].

Limitations. Although prior studies provide insight into the widespread practice of *cookie matching*, they have two significant methodological limitations, which prevent them from observing *all* forms of information sharing between ad exchanges.

1. **Resource Attribution:** These studies cannot determine the precise information flows between ad exchanges, i.e., which parties are sending or receiving information [5]. The fundamental problem is that HTTP requests, and even the DOM tree itself, do not reveal the true sources of resource inclusions in the presence of dynamic code (JavaScript, Flash, *etc.*) from third-parties. For example, a script from `t1.com` embedded in `pub.com` may share identifiers with `t2.com` using dynamic AJAX, but the `Referer` appears to be `pub.com`, thus potentially hiding `t1`'s role as the source of the flow.
2. **Obfuscation:** These studies rely on locating unique user identifiers that are transmitted to multiple third-party domains. Unfortunately, this will miss cases where exchanges send permuted or obfuscated IDs to their partners. Indeed, DoubleClick (the largest ad exchange) is known to do this [3].

In general, these limitations stem from a reliance on analyzing specific *mechanisms* for cookie matching. In this proposal, one of my primary goals is to develop a methodology for detecting cookie matching that is agnostic to the underlying matching mechanism, and instead relies on the fundamental *semantics* of how ad exchanges work.

4 Detecting Information Sharing Between Ad Exchanges

In this section, I propose a novel methodology that is able to detect client- and server-side information flows between arbitrary ad exchanges by leveraging retargeted ads. Our methodology is content- and platform-agnostic and is able to *accurately* attribute the source and destination of the information flow.

4.1 Approach

To study arbitrary information flows in a mechanism-agnostic way, we have to overcome both limitations described in § 3.3.

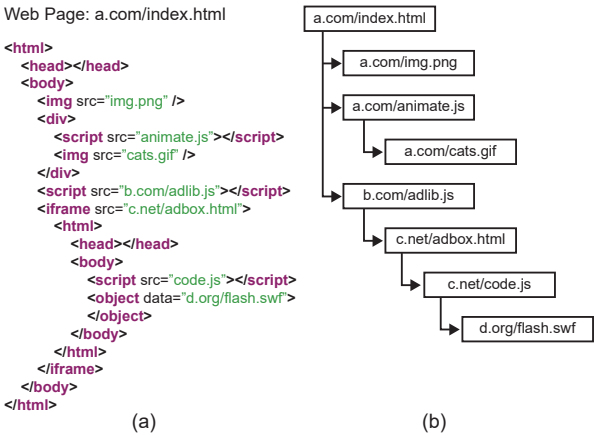

Figure 3: (a) DOM Tree, and (b) Inclusion Tree.

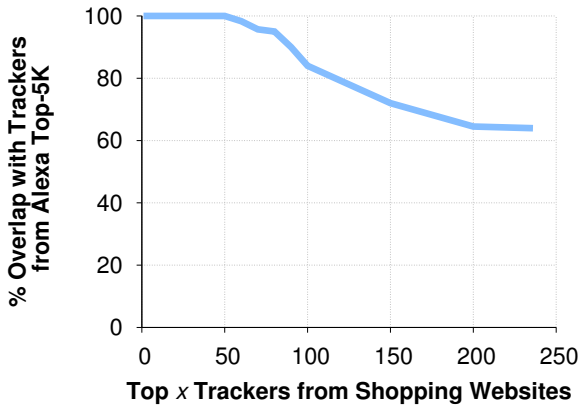

Figure 4: Overlap between frequent trackers on e-commerce sites and Alexa Top-5K sites.

Resource Attribution. For correct resource attribution, we make use of a heavily instrumented version of Chromium [11] that produces *inclusion trees* directly from Chromium's resource loading code. Inclusion trees capture the semantic inclusion structure of resources in a webpage (i.e., which objects cause other objects to be loaded), unlike DOM trees which only capture syntactic structures.

To understand this problem, consider the example DOM tree for `a.com/index.html` in Figure 3(a). Based on the DOM, we might conclude that the chain $a \rightarrow c \rightarrow d$ captures the sequence of inclusions leading from the root of the page to the Flash object from `d.org`.

However, direct use of a webpage's DOM is misleading because the DOM does not reliably record the inclusion relationships between resources in a page. This is due to the ability of JavaScript to manipulate the DOM at run-time, i.e., by adding new inclusions dynamically. As such, while the DOM is a faithful syntactic description of a webpage *at a given point in time*, it cannot be relied upon to extract relationships between included resources. Furthermore, analysis of HTTP request headers does not solve this problem, since the `Referer` is set to the first-party domain even when inclusions are dynamically added by third-party scripts.

Figure 3(b) shows the inclusion tree corresponding to the DOM tree in Figure 3(a). From the inclusion tree, we can see that the true *inclusion chain* leading to the Flash object is $a \rightarrow b \rightarrow c \rightarrow c \rightarrow d$, since the `IFrame` and the Flash are dynamically included by JavaScript from `b.com` and `c.net`, respectively.

Using inclusion chains, we can precisely analyze the provenance of third-party resources included in webpages. The instrumented Chromium accurately captures relationships between elements, regardless of where they are located (e.g., within a single page or across frames) or how the relevant code executes (e.g., via an inline `<script>`, `eval()`, or an event handler).

Obfuscation. We tackle this limitation by relying on a content-agnostic methodology to detect information flows between ad exchanges. Our methodology depends on the following key insight: in most cases, if a user is served a retargeted ad, this proves that ad exchanges shared information about the user (§ 2.2).

To understand this insight, consider that two preconditions must be met for user u to be served a retarget ad for *shop* by DSP d. *First*, either d directly observed u visiting *shop*, or d must be told this information by SSP s. If this condition is not met, then d would not pay the premium price necessary to serve u a retarget. *Second*, if the retarget was served from an ad auction, SSP s and d must be sharing information about u. If this condition is not met, then d would have no way of identifying u as the source of the impression (see § 2.2).

Using this key insight, we reliably infer information flows between SSPs and DSPs, regardless of whether the flow occurs client- or server-side, through the following steps:

- § 4.2: We design *personas* (to borrow terminology from [13] and [19]) that visit specific e-commerce sites in order to elicit retargeted ads from ad exchanges. These sites are carefully chosen to cover different types of products, and include a wide variety of common trackers.
- § 4.3: Using the instumented version of Chromium [11], our personas collect ads by crawling 150 publishers from the Alexa Top-1K. We record *inclusion chains* for all resources encountered during our crawls and use these chains to categorize information flows between ad exchanges.
- § 4.4: Using well-known filtering techniques and crowdsourcing, we identify retargeted ads from our corpus of 571,636 unique crawled images.

If we observe retargeted ads, we know that ad exchanges tracking the user on the *shopper-side* are sharing information with exchanges serving ads on the *publisher-side*.

4.2 Creating Shopper Personas

To signal intent to ad exchanges, we desing 90 *shopping personas*, covering a wide variety of websites. To facilitate this, we leverage the hierarchical categorization of e-commerce sites maintained by Alexa [10]. Our personas cover cover major e-commerce sites (e.g., Amazon and Walmart) and shopping categories (e.g., sports and jewelry). For each persona, we include the top 10 e-commerce sites in the corresponding Alexa category. Furthermore, we manually select 10 product URLs on each of these websites. Thus, each persona visits 100 products URLs. In total, all personas cover 738 unique websites.

Sanity Checking. To make sure that our selected e-commerce sites are embedded with a representative set of trackers, Figure 4 plots the overlap between the trackers we observe on the Alexa Top-5K websites, compared to the top x trackers (i.e., most frequent) we observe on the e-commerce sites. We see that 84% of the top 100 e-commerce trackers are also present in the trackers on Alexa Top-5K sites. These results demonstrate that our shopping personas will be seen by the vast majority of major trackers when they visit our 738 e-commerce sites.

4.3 Ad Collection

To elicit retargeted ads, we selected 150 publishers from Alexa Top-1K websites after manually filtering out those that do not display ads, are non-English, are pornographic, or require

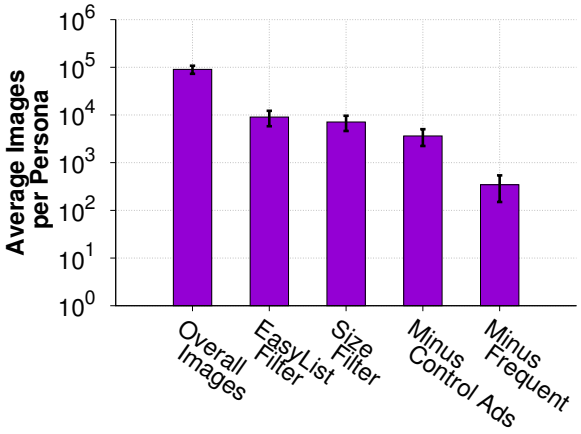

Figure 5: Average number of images per persona, with standard deviation error bars.

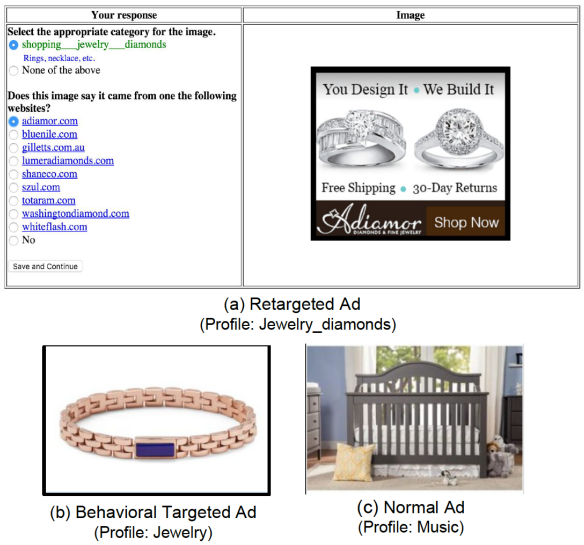

Figure 6: Screenshot of our AMT HIT, and examples of different types of ads.

logging-in to view content (e.g., Facebook). We randomly selected 15 URLs on each publisher to crawl.

We initialized 91 copies of our instrumented Chromium binary: 90 corresponding to our shopper personas, and one which serves as a control. During each *round* of crawling, the personas visit their associated e-commerce sites, then visit the 2,250 publisher URLs (150 publishers $*$ 15 pages per publisher). The control *only* visits the publisher URLs, i.e., it does not browse e-commerce sites, and therefore should never be served retargeted ads. The crawlers are executed in tandem, so they visit the publisher URLs in the same order at the same times.

We conducted nine rounds of crawling between December 4 to 19, 2015. We stopped after 9 rounds because we observed that we only gathered 4% new images during the ninth round. The crawlers recorded inclusion trees, HTTP request and response headers, cookies, and images from all pages.

4.4 Identifying Retargeted Ads

Using our data collection methodology, we collected 571,636 unique images across all crawls. In order to identify ads from these images, we use *EasyList* filter [28] provided by *AdBlock-Plus* [6]. In our case, we look at the inclusion chain for each image, and filter out those in which none of the URLs in the chain are a hit against EasyList. This reduces the set to 93,726 unique images. We also filter out all images with dimensions $< 50 \times 50$ pixels. These images are too small to be ads; most are 1×1 tracking pixels.

Since we are looking for retargeted ads, they should, by definition, only appear to personas that visit a specific e-commerce site. In other words, any ad that was shown to our control account (which visits no e-commerce sites) is either untargeted or contextually targeted, and can be discarded. Furthermore, any ad shown to >1 persona may be behaviorally targeted,

but it cannot be a retarget, and is therefor filtered out. After applying this filter, we are left with 31,850 ad images. Figure 5 shows the average number of images remaining per persona after applying various filters.

Crowdsourcing. Even after applying all filters, we cannot be sure whether an ad is a retargeted ad or just a contextual ad. To clearly distinguish retargeted ads, we have to manually go over the 31,850 images. However, given the large number of ads in our corpus, we decided to crowdsource labels from workers on Amazon Mechanical Turk (AMT). We constructed Human Intelligence Tasks (HITs) that ask workers to label 30 ads, 27 of which are unlabeled, and 3 of which are known to be retargeted ads and serve as controls (we manually identified 1,016 retargets from our corpus of 31,850 to serve as these controls).

Figure 6(a) shows a screenshot of our HIT. On the right is an ad image, and on the left we ask the worker two questions:

1. Does the image belong to one of the following categories (with "None of the above" being one option)?
2. Does the image say it came from one of the following websites (with "No" being one option)?

The purpose of question (1) is to isolate behavioral and retargeted ads from contextual and untargeted ads (e.g., Figure 6(c), which was served to our *Music* persona). The list for question (1) is populated with the shopping categories associated with the persona that crawled the ad. For example, as shown in Figure 6(a), the category list includes "shopping_jewelry_diamonds" for ads shown to our *Diamond Jewelry* persona. In most cases, this list contains exactly one entry, although there are rare cases where up to 3 categories are in the list.

If the worker does not select "None" for question (1), then they are shown question (2). Question (2) is designed to separate retargets from behavioral targeted ads. The list of websites for question (2) is populated with the e-commerce sites visited by the persona that crawled the ad. For example, in Figure 6(a), the ad clearly says "Adiamor", and one of the sites visited by the persona is `adiamor.com`; thus, this image is likely to be a retarget.

To make sure that the quality of labeling is good, we restrict our HITs to workers that have completed $\geq$50 HITs and have an approval rating $\geq$95%. Additionally, we reject a HIT if the worker mislabels $\geq$2 of the control images (i.e., known retargeted ads) and get new labels for rejected HITs by another worker. We obtain two labels on each unlabeled image by different workers. For 92.4% of images both labels match, so we accept them. We manually labeled the divergent images ourselves to break the tie.

The workers from AMT successfully identified 1,359 retargeted ads. However, it is possible that they failed to identify some retargets, i.e., there are false negatives. This may occur in cases like Figure 6(b): it is not clear if this ad was served as a behavioral target based on the persona's interest in jewelry, or as a retarget for a specific jeweler.

To mitigate this issue, we manually examined all 7,563 images that were labeled as behavioral ads by the workers. In addition to the images themselves, we also looked at the inclusion chains for each image. In many cases, the URLs reveal that specific e-commerce sites visited by our personas hosted the images, indicating that the ads are retargets. For example, Figure 6(b) is actually part of a retargeted ad from `fossil.com`. Our manual

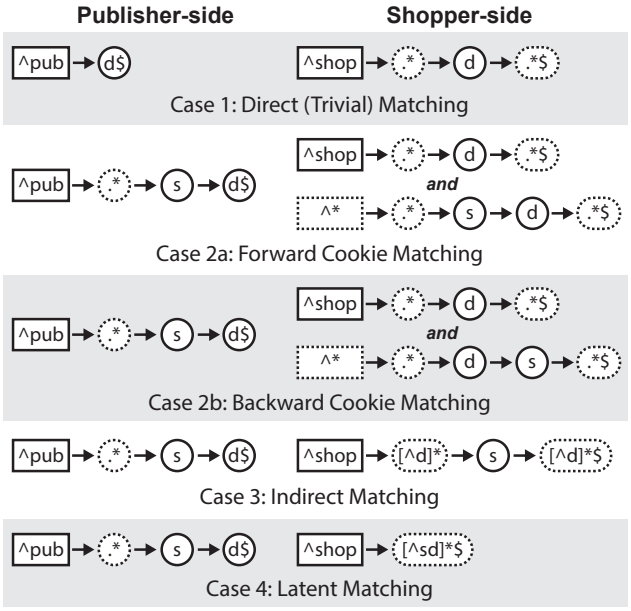

Figure 7: Regex-like rules we use to identify different types of ad exchange interactions. *shop* and *pub* refer to chains that begin at an e-commerce site or publisher, respectively. d is the DSP that serves a retarget; s is the predecessor to d in the publisher-side chain, and is most likely an SSP holding an auction. Dot star ($.*$) matches any domains zero or more times.

analysis uncovered an additional 3,743 retargeted ads, bringing the total to 5,102. These retargets advertise 281 distinct e-commerce websites (38% of all e-commerce sites).

4.5 Information Flow Categorization

Now that we have identified 5,102 retargeted ads, our goal is to determine information flows between ad exchanges by analyzing their associated inclusion chains. Specifically, we seek to answer two fundamental questions: *who* is sharing user data, and *how* does the sharing take place (e.g., client-side via cookie matching, or server-side)? To answer these questions, we categorize the 35,448 *publisher-side* inclusion chains corresponding to the 5,102 retargeted ads in our data. Note that 1) we observe some retargeted ads multiple times, resulting in multiple chains, and 2) the chains for a given unique ad may not be identical.

Terminology. Each retargeted ad was served to our persona via a *publisher-side* chain. *pub* is the domain of the publisher at the root of the chain, while d is the domain at the end of the chain that served the ad. Typically, d is a DSP. If the retarget was served via an auction, then an SSP s must immediately precede d in the publisher-side chain. Each retarget advertises a particular e-commerce site. *shop* is the domain of the e-commerce site corresponding to a particular retargeted ad. To categorize a given publisher-side chain, we must also consider the corresponding *shopper-side* chains rooted at *shop*.

4.5.1 Categorization Rules

Case 1: Direct Matches. These matches are the simplest type of chains that can be used to serve a retargeted ad and require no information sharing between ad exchanges. As shown in Figure 7, for us to categorize a publisher-side chain as a direct match, it must be exactly length two, with a direct resource inclusion request from *pub* to d. d receives any cookies they have stored on the persona inside this request, and thus it is trivial for d to identify our persona.

On the shopper-side, the only requirement is that d observed our persona browsing *shop*. If d does not observe our persona at *shop*, then d would not serve the persona a retargeted ad for *shop*. d is able to set a cookie on our persona, allowing d to re-identify the persona in future.

Case 2: Cookie Matching. As the name implies, chains in this category correspond to instance where an auction is held on the publisher-side, and we observe direct resource inclusion requests between the SSP and DSP, implying that they are matching cookies.

As shown in Figure 7, for us to categorize a publisher-side chain as cookie matching, s and d must be adjacent at the end of the chain. On the shopper-side, d must observe the persona at *shop*. Lastly, we must observe a request from s to d or from d to s in some chain before the retargeted ad is served. These requests capture the moment when the two ad exchanges match their cookies. Many pairs of ad exchanges engage in both *forward* ($s \rightarrow d$) and *backward* ($d \rightarrow s$) cookie matches.

Case 3: Indirect Matching. This sort of matching occurs when an SSP sends meta-data about a user to a DSP, to help them determine if they should bid on an impression. With respect to retargeted ads, the SSP tells the DSPs about the browsing history of the user, thus enabling the DSPs to serve retargets for specific retailers, even if the DSP never directly observed the user browsing the retailer (hence the name, *indirect*). Note that no cookie matching is necessary in this case for DSPs to serve retargeted ads.

As shown in Figure 7, the crucial difference between cookie matching chains and indirect chains is that d *never* observes our persona at *shop*; only s observes our persona at *shop*. Thus, by inductive reasoning, we must conclude that s shares information about our persona with d, otherwise d would never serve the persona a retarget for *shop*.

Case 4: Latent Matching. As shown in Figure 7, the defining characteristic of latent chains is that neither s nor d observe our persona at *shop*. This begs the question: how do s and d know to serve a retargeted ad for *shop* if they never observe our persona at *shop*? The most reasonable explanation is that some other ad exchange x that is present in the shopper-side chains shares this information with d behind-the-scenes.

We hypothesize that the simplest way for ad exchanges to implement latent matching is by having x and d share the same unique identifiers for users. Although x and d are different domains, and are thus prevented by the SOP from reading each others' cookies, both ad exchanges may use the same deterministic algorithm for generating user IDs (e.g., by relying on IP addresses or browser fingerprints). However, as we will show, these synchronized identifiers are not necessarily visible from the client-side (i.e., the values of cookies set by x and d may be obfuscated), which prevents trivial identification of latent cookie matching.

Type	Unclustered Chains	%	Clustered Chains	%
Direct	1770	5%	8449	24%
Forward Cookie Match	24575	69%	25873	73%
Backward Cookie Match	19388	55%	24994	70%
Indirect Match	2492	7%	178	1%
Latent Match	5362	15%	343	1%
No Match	775	2%	183	1%

Table 1: Results of categorizing publisher-side chains, before and after clustering domains.

4.6 Results

We applied the rules in Figure 7 to all 35,448 publisher-side chains in our dataset twice. First, we categorized the raw, unmodified chains; then we *clustered* domains that belong to the same companies, and categorized the chains again. For example, Google owns `youtube.com`, `doubleclick.com`, and `2mdn.net`; in the clustered experiments, we replace all instances of these domains with `google.com`.

Table 1 presents the results of our categorization. The first thing we observe is that cookie matching is the most frequent classification by a large margin. This conforms to our expectations, given that RTB is widespread in today's ad ecosystem, and major exchanges like DoubleClick support it [25].

The next interesting feature that we observe is that indirect and latent matches are relatively rare (7% and 15%, respectively). Again, this is expected, since these types of matching are more exotic and require a greater degree of collaboration between ad exchanges to implement. Furthermore, the percentage of indirect and latent matches drops to 1% when we cluster domains. To understand why this occurs, consider the following real-world example chains:

Publisher-side: $pub \rightarrow rubicon \rightarrow googlesyndication$

Shopper-side: $shop \rightarrow doubleclick$

According to the rules in Figure 7, this appears to be a latent match, since Rubicon and Google Syndication do not observe our persona on the shopper-side. However, after clustering the Google domains, this will be classified as cookie matching (assuming that there exists at least one other request from Rubicon to Google).

The above example is extremely common in our dataset: 731 indirect chains become cookie matching chains after we cluster the Google domains *alone*. Importantly, this finding provides strong evidence that Google does in fact use latent matching to share user tracking data between its various domains. Although this is allowed in Google's terms of service as of 2014 [39], our results provide direct evidence of this data sharing with respect to serving targeted ads.

The final takeaway from Table 1 is that the number of uncategorized chains that do not match any of our rules is extremely low (1-2%). These publisher-side chains are likely to be false positives, i.e., ads that are not actually retargeted. These results suggest that our image labeling approach is very robust, since the vast majority of chains are properly classified as direct or cookie matches.

Participant 1		Participant 2	Chains	Ads	Heuristics
criteo	↔	googlesyndication	9090	1887	↔: US
criteo	↔	doubleclick	3610	1144	→: E, US ←: DC, US
criteo	↔	rubiconproject	1586	749	↔: E, US
mythings	↔	mythingsmedia	478	52	→: E, US ←: US
criteo	↔	bidswitch	112	78	→: E, US ←: US
rubiconproject	↔	steelhousemedia	86	30	↔: E
googlesyndication	↔	steelhousemedia	47	22	-
adtechus	→	adacado	36	18	-
googlesyndication	↔	2mdn	40	19	→: US ←: -
googlesyndication	↔	adlegend	31	22	-

Table 2: 10 sampled cookie matching partners in our dataset. The arrow signifies whether we observe forward matches (→), backward matches (←), or both (↔). The heuristics for detecting cookie matching are: *DC* (match using DoubleClick URL parameters), *E* (string match for exact cookie values), *US* (URLs that include parameters like "usersync"), and - (no identifiable mechanisms).

The results from Table 1 confirm that cookie matching is ubiquitous on today's web, and that this information sharing fuels highly targeted advertisements.

Who Is Cookie Matching? Table 2 shows a sample of 10 popular pairs of domains that we observe matching cookies. The arrows indicate the direction of matching (forward, backward, or both). "Ads" is the number of unique retargets served by the pair, while "Chains" is the total number of associated publisher-side chains. In total we observe 200 cookie matching partners.

We observe that cookie matching frequency is heavily skewed towards several heavy-hitters. In aggregate, Google's domains are most common, which makes sense given that Google is the largest ad exchange on the Web today. The second most common is Criteo; this result also makes sense, given that Criteo specializes in retargeted advertising [23].

Interestingly, we observe a great deal of heterogeneity with respect to the directionality of cookie matching. Some boutique exchanges, like Adacado, only ingest cookies from other exchanges. Others, like Criteo, are omnivorous, sending or receiving data from any and all willing partners. These results suggest that some participants are more wary about releasing their user identifiers to other exchanges.

Comparison With Prior Work. To understand what fraction of cookie matching relationships will be missed by the heuristic detection approaches used by prior work [58, 5, 72, 35], we applied the three of them to our dataset. Specifically, for each pair (s, d) of exchanges that we categorize as cookie matching, we apply the following tests to the HTTP headers of requests between s and d or vice-versa:

1. We look for specific keys that are known to be used by DoubleClick and other Google domains for cookie matching (e.g., "google_nid" [72]).
2. We look for cases where unique cookie values set by one participant are included in requests sent to the other participant.
3. We look for keys with revealing names like "usersync" that frequently appear in requests between participants in our data.

As shown in the "Heuristics" column in Table 2, in the majority of cases, heuristics are able to identify cookie matching between the participants. Interestingly, we observe that

the mechanisms used by some pairs (e.g., Criteo and DoubleClick) change depending on the directionality of the cookie match, revealing that the two sides have different cookie matching APIs.

However, for 31% of our cookie matching partners, the heuristics are unable to detect signs of cookie matching. We hypothesize that this is due to obfuscation techniques employed by specific ad exchanges. In total, there are 4.1% cookie matching chains that would be completely missed by heuristic tests. This finding highlights the limitations of prior work, and bolsters the case for our mechanism-agnostic classification methodology.

5 Cross-Device Tracking

The rise of mobile devices such as smartphones and tablets has significantly altered how users access the internet. According to a recent study, 77% of Americans now own a smartphone [21], and more than 40% use multiple devices to access the internet [30]. This change in user behavior has also forced A&A companies to evolve. Online tracking is not limited to desktop devices anymore: A&A companies have come up with various ways to track users on the mobile devices as well [87, 15, 31, 40, 78, 76].

However, to gain a complete picture of user behavior and interests across all devices, A&A companies need to identify all devices associated with a particular user. The process of tracking users across multiple devices is known as cross-device tracking. Although we have seen in § 3 that there has been a significant amount of research on understanding the state of tracking on desktop and mobile devices separately, there is little work on the cross-device tracking ecosystem. While two studies [16, 92] have highlighted the potential for and existence of cross-device tracking, they have not shed light on the underlying mechanics of this process. For example, these studies do not tell us which attribute(s) (e.g., email address, username, advertising ID) facilitated cross-device tracking, nor do they identify the A&A companies which share information across multiple devices.

The detection of information sharing between A&A companies across devices is particularly challenging, since the tracking mechanisms on these devices differ. For example, cookies are used to track users on desktop devices, while advertising ID is used to track users on mobile apps. Since prior work [58, 5, 72, 35] looks for specific, repeated identifiers to detect cookie matches on desktop, it cannot be used to detect information flows between A&A companies across multiple devices.

To this end, I propose to study cross-device tracking using the methodology I developed in § 4. Since my methodology does not look for patterns or identifiers in the network traffic, but rather relies on the causal inference of how a retargeted ad is shown, I can use it to study cross-device tracking. In particular, I propose to investigate the following questions:

1. Which A&A company observes the user on device d_1, and which A&A company shows the advertisement on device d_2? If the two companies are different, then it implies information flow between them.
2. *Which* information was shared between the A&A companies on the two devices to link the user? Is it a deterministic identifier (e.g., (e.g., email address, advertsing ID, IMEI) or a probabilistic one (e.g., location, Wifi network, browsing behavior)?

3. *How* prevalent is cross-device tracking and which A&A companies are involved?

Methodology. I plan to use the same high-level methodology as the one used in § 4. Instead of creating personas and collecting ads on the same device, we create a persona and browse products (e.g., running shoes) on device d_1 (e.g., Android smartphone), and then visit publishers to collect ads on device d_2 (e.g., desktop). If we observe retargeted ads on d_2, then it means that the user has been tracked across the two devices. We can use the inclusion chains captured by our instrumented version of Chrome to analyze which A&A companies were involved in tracking the user across devices. We have successfully ported our Chrome instrumentation to Android by hooking the WebView interface (which, fortunately, uses a Chrome behind the scenes).

Figuring out the exact identifier that was used to track the user across the two devices is challenging. To isolate the identifier that gave away the user, we have to conduct controlled experiments. For example, if we want to know whether email address e_1 is used to track users across devices, we would have to explicitly sign in with e_1 to mobile apps on d_1 to browse products. In addition to this, we have to make sure that *only* email address is leaked while browsing, and other identifiers are either blocked (e.g., location) or randomized (e.g., advertising ID). We achieve this level of control by instrumenting Chrome on the desktop side, and using the Xposed framework on Android. Xposed modules allow us to spoof stateful identifiers on Android, such as advertising ID, IMEI, and MAC address.

Once we have done that, we would have to visit publishers on d_2 that explicitly request an email address from the user. We would use e_1 to sign-in to the publisher or to subscribe to promotional emails. Finally, if we observe retargeted ads on d_2, we can infer that email address e_1 enabled cross-device tracking.

Using this methodology, we can conduct large-scale experiments. We can control the identifier we want to leak to A&A companies during the browsing period, and then check for the presence of retargeted ads to verify the usage of that particular identifier for cross-device tracking. I plan on investigating both deterministic and probabilistic identifiers for this study.

This project is currently active and is in early stages. I have completed experiments to check the usage of IP address for cross-device tracking. I aim to wrap up this project in early 2019.

6 Information Diffusion in the Advertising Graph

As we note in § 2, the online advertising ecosystem has become enormously complex. The rise of RTB has caused ad networks to shift towards ad exchanges, and caused advertising companies to specialize into particular roles like DSPs or SSPs. The information flows between these entities encode important information about how personal data moves through the online advertising ecosystem.

A natural way to model this complex ecosystem is in the form of a graph. Graph models that accurately capture the relationships between publishers and advertising companies are extremely important for practical applications, such as estimating revenue of advertising companies [37] or predicting whether a given domain is a tracker [46]. Keeping in mind

the rise of RTB and the close collaboration among A&A companies due to *cookie matching* (§ 4.6), we can capture diffusion of tracking data in the advertising ecosystem through graph modeling.

However, to date, technical limitations have prevented researchers from developing accurate graph models of the online advertising ecosystem. For example, Gomer et al. [38] propose a *Referer* graph, where nodes represent publishers or advertising domains, and two nodes a_i and a_j are connected if an HTTP message to a_j is observed with a_i as the HTTP `Referer`. Unfortunately, as we note in § 3.3, dynamic inclusions by JavaScript can lead to mis-attributed `Referer`. This causes erroneous edges in cases where a third-party script is embedded directly into a first-party context (i.e., is not sandboxed in an `iframe`). Furthermore, there can be mis-attributed or missing edges due to the inability of prior work (§ 3.3) to capture *all* information flows (*cookie matches*) accurately.

Since the methodology I proposed in § 4 can *accurately* detect information flows between A&A companies, I propose a novel and accurate representation of the advertising graph called an *Inclusion* graph, using the inclusion chains collected through extensive experiments in § 4.5.

In the *Inclusion* graph, nodes represent Advertising and Analytics (A&A) companies, publishers, or other online services. Edges capture the relationships between these actors, such as resource inclusion or information flow (cookie matching). This representation is a directed graph of publishers and A&A companies. An edge $d_i \rightarrow d_j$ exists if we have ever observed domain d_i including a resource from d_j. The weight of an edge $d_i \rightarrow d_j$ encodes the number of times a resource from d_i sends an HTTP request to d_j. Edges may exist from publisher to A&A nodes, or between A&A nodes.

Taking RTB into account, we can use the *Inclusion* graph to model the diffusion of user tracking data across the advertising ecosystem. By simulating the browsing behavior using the methodology from Burklen et al. [17], we generate browsing traces for 200 users. Each user, on average, generated 5,343 impressions (page visits) across 190 publishers.

Through simulations, we show that the *Inclusion* graph can be used to implement empirically-driven simulations of the online ad ecosystem. Our results demonstrate that due to the involvement in RTB, top A&A companies observe the vast majority of users' browsing history. Even under realistic conditions where only a small number (37) of well-connected ad exchanges indirectly share impressions during RTB auctions, the top 10% of A&A companies observe more than 91% impressions (page visits) and 82% visited publishers.

We also evaluate a variety of ad and tracker blocking strategies [26, 6, 36] in the context of our models, to understand their effectiveness at stopping A&A companies from learning users' browsing data. We observe that AdBlock Plus [6] has essentially zero impact on the fraction of impressions observed by A&A companies. The problem is that the major ad exchanges are all present in the Acceptable Ads whitelist [7], and thus all of their partners are also able to observe the impressions during RTB, even if they are (sometimes) prevented from actually showing ads to the user. Disconnect [26] and Ghostery [36] both perform better than AdBlockPlus, with blocking upto 50% and 25% impression flows to the top 10 A&A companies. We further observe that even when strong blocking is used, top A&A domains can still observe anywhere from 40–70% of simulated users' impressions.

This project is currently under submission at *The Privacy Enhancing Technologies Symposium* (PETS).

7 Towards Automation of Privacy Protection Filter Lists

Users have grown increasingly concerned regarding the amounts and types of data collected about them via online tracking [62, 85]. This has led to the proliferation of tools and techniques to block online ads and prevent tracking. Measurement studies estimate that Adblock Plus (the most popular ad blocking browser extension) is used by roughly 16–37% of web users [75, 60], and numerous other extensions like Ghostery, Disconnect, Privacy Badger, and uBlock Origin have devoted user bases.

To block ads and trackers, these tools rely on crowdsourced filter lists such as EasyList [28] and EasyPrivacy [29]. Over the years, crowdsourced contributors have added regular expression like rules to these lists. Adblocking extensions use these rules to determine whether a requested URL[1] should be blocked or not. Despite being popular, these filter lists have certain limitations:

- These lists have grown in size over the years. As of this writing, EasyList has roughly 73K entries. This can cause web pages to have a larger load time.
- These lists are not easy to manage since filter rules can be complex. A generic rule can cause site breakage. To counter this, filter lists add exceptions, which in turn can be quite complicated. Furthermore, these lists can suffer from accuracy on blocking trackers on less popular websites [32].

I propose to automate the creation of blocking filter list by using Authorized Digital Sellers (*ADS.txt*) protocol [8]. This protocol was introduced by Interactive Advertising Bureau (IAB) in 2017, which allows publishers to specify which SSPs or exchanges can sell their inventory. Publishers can do so by listing authorized sellers in a file, called `ads.txt`, hosted at the root of their website. The onus is on the DSPs participating in RTB auctions to crawl the `ads.txt` for a specific publisher to make sure that the bid request is coming from an exchange which is authorized by the publisher to sell its inventory. *ADS.txt* adoption is on the rise; DoubleClick (the largest ad exchange) has announced that it will only buy inventory from sources identified as authorized sellers by the publisher [27].

We have been crawling `ads.txt` files for top 100K Alexa websites every 15 days since January 2018. Additionally, we also collect inclusion chains from these publishers, right after the `ads.txt` crawl. We plan to combine all the inclusion chains with authorized sellers, to form a graph. Then, through label propagation, we plan to identify nodes which should make on the filter list.

Following are some of the key research questions to tackle:

1. Is the coverage of *ADS.txt* good enough to create an effective filter list?
2. How does this filter list fare with current filter lists like EasyList and EasyPrivacy?

We have been collecting data for this project since January 2018. I aim to wrap up this project by April 2019.

[1] Some extensions like Disconnect block entire domains rather than URLs.

8 Research Plan

Table 3 presents my plan for completing the research outlined in this proposal.

Timeline	**Work**	**Progress**
Usenix Security 2016	Detecting information flows between ad exchanges	Published
PETS 2018	Modeling advertising graph using inclusion chains	Published
Present - APR 2019	Automating adblocking filter list	Ongoing
Present - MAY 2019	Cross-device tracking	Ongoing
APR - JUL 2019	Thesis writing	-
AUG 2019	Thesis defense	-

Table 3: Plan for completion of my research.

References

[1] Real-time bidding in the united states and worldwide. Karsten Weide, October 2012. http://vincenttessier.fr/wp-content/uploads/2012/11/IDC-RTB-research-2012.pdf.

[2] Share of real-time bidding in digital display advertising spending in the united states from 2012 to 2018. Statista, March 2014. https://www.statista.com/statistics/267762/share-of-rtb-in-digital-display-ad-spend-in-the-us/.

[3] Real-time bidding protocol, February 2016. https://developers.google.com/ad-exchange/rtb/cookie-guide.

[4] U.s ad spending: The emarketer forecast for 2017. eMarketer, March 2017. https://www.emarketer.com/Report/US-Ad-Spending-eMarketer-Forecast-2017/2001998.

[5] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proc. of CCS*, 2014.

[6] Adblock plus: Surf the web without annoying ads! eyeo GmbH. https://adblockplus.org.

[7] Allowing acceptable ads in adblock plus. eyeo GmbH. https://adblockplus.org/acceptable-ads.

[8] Ads.txt âĂŞ authorized digital sellers. IAB Tech Lab. https://iabtechlab.com/ads-txt/.

[9] Lalit Agarwal, Nisheeth Shrivastava, Sharad Jaiswal, and Saurabh Panjwani. Do not embarrass: Re-examining user concerns for online tracking and advertising. In *Proc. of the Workshop on Usable Security*, 2013.

[10] Alexa. The top 500 sites on th web. Alexa. https://www.alexa.com/topsites/category/Top/Shopping.

[11] Sajjad Arshad, Amin Kharraz, and William Robertson. Include me out: In-browser detection of malicious third-party content inclusions. In *Proc. of Intl. Conf. on Financial Cryptography*, 2016.

[12] Mika Ayenson, Dietrich James Wambach, Ashkan Soltani, Nathan Good, and Chris Jay Hoofnagle. Flash cookies and privacy ii: Now with html5 and etag respawning. *Available at SSRN 1898390*, 2011.

[13] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma, and S. Muthukrishnan. Adscape: Harvesting and analyzing online display ads. In *Proc. of WWW*, 2014.

[14] Muhammad Ahmad Bashir, Sajjad Arshad, and Christo Wilson. "Recommended For You": A First Look at Content Recommendation Networks. In *Proc. of IMC*, 2016.

[15] Theodore Book and Dan S. Wallach. A case of collusion: A study of the interface between ad libraries and their apps. In *Proceedings of the Third ACM Workshop on Security and Privacy in Smartphones*, SPSM '13, 2013.

[16] Justin Brookman, Phoebe Rouge, Aaron Alva, and Christina Yeung. Cross-device tracking: Measurement and disclosures. In *Proc. of PETS*, 2017.

[17] Susanne Burklen, Pedro Jose Marron, Serena Fritsch, and Kurt Rothermel. User centric walk: An integrated approach for modeling the browsing behavior of users on the web. In *Annual Symposium on Simulation*, April 2005.

[18] Aaron Cahn, Scott Alfeld, Paul Barford, and S. Muthukrishnan. An empirical study of web cookies. In *Proc. of WWW*, 2016.

[19] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. I always feel like somebody's watching me: Measuring online behavioural advertising. In *Proc. of ACM CoNEXT*, 2015.

[20] Claude Castelluccia, Mohamed-Ali Kaafar, and Minh-Dung Tran. Betrayed by your ads!: Reconstructing user profiles from targeted ads. In *Proc. of PETS*, 2012.

[21] Pew Research Center. Mobile fact sheet. pewinternet.org, February 2018. http://www.pewinternet.org/fact-sheet/mobile/.

[22] Farah Chanchary and Sonia Chiasson. User perceptions of sharing, advertising, and tracking. In *Proc. of the Workshop on Usable Security*, 2015.

[23] Criteo ranking by Econsultancy. http://www.criteo.com/resources/e-consultancy-display-retargeting-buyers-guide/.

[24] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. In *Proc. of PETS*, 2015.

[25] Double Click RTB explained. https://developers.google.com/ad-exchange/rtb/.

[26] Disconnect defends the digital you. Disconnect Inc. https://disconnect.me/.

[27] Google announces new anti-fraud initiatives for doubleclick bid manager. Martech Today, September 2017. https://martechtoday.com/google-doubleclick-new-ad-fraud-measures-204475.

[28] Easylist. The EasyList authors. https://easylist.to/easylist/easylist.txt.

[29] Easyprivacy. The EasyList authors. https://easylist.to/easylist/easyprivacy.txt.

[30] Econsultancy. More than 40 econsultancy.com, March 2014. https://www.econsultancy.com/blog/64464-more-than-40-of-online-adults-are-multi-device-users-stats.

[31] Manuel Egele, Christopher Kruegel, Engin Kirda, and Giovanni Vigna. Pios: Detecting privacy leaks in ios applications. In *Proc of NDSS*, 2011.

[32] Steven Englehardt and Arvind Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proc. of CCS*, 2016.

[33] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. Cookies that give you away: The surveillance implications of web tracking. In *Proc. of WWW*, 2015.

[34] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In *Proc of. Traffic Monitoring and Analysis*, 2014.

[35] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Tracking personal identifiers across the web. In *Proc. of PAM*, 2016.

[36] Ghostery: faster, cleaner, and safer browsing. Cliqz International GmbH i.Gr. https://www.ghostery.com/.

[37] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *Proc. of IMC*, 2013.

[38] R. Gomer, E. M. Rodrigues, N. Milic-Frayling, and M. C. Schraefel. Network analysis of third party tracking: User exposure to tracking cookies through search. In *Prof. of IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2013.

[39] Gloria Goodale. Privacy concerns? what google now says it can do with your data. Christian Science Monitor, April 2014. http://www.csmonitor.com/USA/2014/0416/Privacy-concerns-What-Google-now-says-it-can-do-with-your-data-video.

[40] Michael C. Grace, Wu Zhou, Xuxian Jiang, and Ahmad-Reza Sadeghi. Unsafe exposure analysis of mobile in-app advertisements. In *Proceedings of the Fifth ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WISEC '12, 2012.

[41] Saikat Guha, Bin Cheng, and Paul Francis. Challenges in measuring online advertising systems. In *Proc. of IMC*, 2010.

[42] Raymond Hill. ws-gateway websocket circumvention ? #1936. GitHub, August 2016. https://github.com/gorhill/uBlock/issues/1936.

[43] David Howell. How to protect your privacy and remove data from online services. Tech Radar, January 2015. http://www.techradar.com/news/internet/how-to-protect-your-privacy-and-remove-data-from-online-services-1291515.

[44] Muhammad Ikram, Hassan Jameel Asghar, Mohamed Ali Kâafar, Balachander Krishnamurthy, and Anirban Mahanti. Towards seamless tracking-free web: Improved detection of trackers via one-class learning. *PoPETs*, 2017(1):79–99, 2017.

[45] Sakshi Jain, Mobin Javed, and Vern Paxson. Towards mining latent client identifiers from network traffic. *PoPETs*, 2016(2):100–114, 2016.

[46] Vasiliki Kalavri, Jeremy Blackburn, Matteo Varvello, and Konstantina Papagiannaki. Like a pack of wolves: Community structure of web trackers. In *Proc. of Passive and Active Measurement*, 2016.

[47] Samy Kamkar. Evercookie - virtually irrevocable persistent cookies., September 2010. http://samy.pl/evercookie/.

[48] T. Kohno, A. Broido, and K. Claffy. Remote physical device fingerprinting. *IEEE Transactions on Dependable and Secure Computing*, 2(2):93–108, 2005.

[49] Balachander Krishnamurthy, Delfina Malandrino, and Craig E. Wills. Measuring privacy loss and the impact of privacy protection in web browsing. In *Proc. of the Workshop on Usable Security*, 2007.

[50] Balachander Krishnamurthy, Konstantin Naryshkin, and Craig Wills. Privacy diffusion on the web: A longitudinal perspective. In *Proc. of WWW*, 2009.

[51] Balachander Krishnamurthy and Craig Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Proc. of W2SP*, 2011.

[52] Balachander Krishnamurthy and Craig E. Wills. Generating a privacy footprint on the internet. In *Proc. of IMC*, 2006.

[53] Mathias Lécuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. Xray: Enhancing the web's transparency with differential correlation. In *Proc. of USENIX Security Symposium*, 2014.

[54] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proc. of CCS*, 2015.

[55] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. What matters to users?: Factors that affect users' willingness to share information with online advertisers. In *Proc. of the Workshop on Usable Security*, 2013.

[56] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *Proc. of USENIX Security Symposium*, Austin, TX, 2016.

[57] Tai-Ching Li, Huy Hang, Michalis Faloutsos, and Petros Efstathopoulos. Trackadvisor: Taking back browsing privacy from third-party trackers. In *Proc. of PAM*, 2015.

[58] Bin Liu, Anmol Sheth, Udi Weinsberg, Jaideep Chandrashekar, and Ramesh Govindan. Adreveal: Improving transparency into online targeted advertising. In *Proc. of HotNets*, 2013.

[59] Miguel Malheiros, Charlene Jennett, Snehalee Patel, Sacha Brostoff, and Martina Angela Sasse. Too close for comfort: A study of the effectiveness and acceptability of rich-media personalized advertising. 2012.

[60] Matthew Malloy, Mark McNamara, Aaron Cahn, and Paul Barford. Ad blockers: Global prevalence and impact. In *Proc. of IMC*, 2016.

[61] Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *Proc. of IEEE Symposium on Security and Privacy*, 2012.

[62] Aleecia M. McDonald and Lorrie Faith Cranor. Americans' attitudes about internet behavioral advertising practices. In *Proc. of WPES*, 2010.

[63] MediaMath. Strength in numbers., 2007. `http://www.mediamath.com`.

[64] Georg Merzdovnik, Markus Huber, Damjan Buhov, Nick Nikiforakis, Sebastian Neuner, Martin Schmiedecker, and Edgar R. Weippl. Block me if you can: A large-scale study of tracker-blocking tools. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, 2017.

[65] Keaton Mowery and Hovav Shacham. Pixel perfect: Fingerprinting canvas in html5. In *Proc. of W2SP*, 2012.

[66] Mozilla. Same-origin policy., May 2008. `https://developer.mozilla.org/en-US/docs/Web/Security/Same-origin_policy`.

[67] Muhammad Haris Mughees, Zhiyun Qian, and Zubair Shafiq. Detecting anti ad-blockers in the wild. *PoPETs*, 2017(3):130, 2017.

[68] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. Privaricator: Deceiving fingerprinters with little white lies. In *Proc. of WWW*, 2015.

[69] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proc. of IEEE Symposium on Security and Privacy*, 2013.

[70] Rishab Nithyanand, Sheharbano Khattak, Mobin Javed, Narseo Vallina-Rodriguez, Marjan Falahrastegar, Julia E. Powles, Emiliano De Cristofaro, Hamed Haddadi, and Steven J. Murdoch. Adblocking and counter blocking: A slice of the arms race. In *Proc. of FOCI*, 2016.

[71] Lukasz Olejnik, Claude Castelluccia, and Artur Janc. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. In *Proc. of HotPETs*, 2012.

[72] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. Selling off privacy at auction. In *Proc of NDSS*, 2014.

[73] OpenX. Programmatic advertising., 2007. `https://www.openx.com`.

[74] Fotios Papaodyssefs, Costas Iordanou, Jeremy Blackburn, Nikolaos Laoutaris, and Konstantina Papagiannaki. Web identity translator: Behavioral advertising and identity privacy with wit. In *Proc. of HotNets*, 2015.

[75] Enric Pujol, Oliver Hohlfeld, and Anja Feldmann. Annoyed users: Ads and ad-block usage in the wild. In *Proc. of IMC*, 2015.

[76] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. Apps, trackers, privacy and regulators: A global study of the mobile tracking ecosystem. In *Proc of NDSS*, 2018.

[77] Maire Reavy. Webrtc privacy. mozillamediagoddess.org, September 2015. `https://mozillamediagoddess.org/2015/09/10/webrtc-privacy/`.

[78] Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout, and David Choffnes. Recon: Revealing and controlling pii leaks in mobile network traffic. In *Proc. of MobiSys*, 2016.

[79] Technobuffalo.com. EasyList Forum, July 2016. `https://forums.lanik.us/viewtopic.php?p=110902`.

[80] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *Proc. of NSDI*, 2012.

[81] RubiconProject. The global exchnage for advertising., 2007. `https://rubiconproject.com`.

[82] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. Flash cookies and privacy. In *AAAI Spring Symposium: Intelligent Information Privacy Management*, 2010.

[83] Lincoln Spector. Online privacy tips: 3 ways to control your digital footprint. PC World, January 2016. `http://www.pcworld.com/article/3020163/internet/online-privacy-tips-3-ways-to-control-your-digital-footprint.html`.

[84] Oleksii Starov and Nick Nikiforakis. Extended tracking powers: Measuring the privacy diffusion enabled by browser extensions. In *Proc. of WWW*, 2017.

[85] Joseph Turow, Michael Hennessy, and Nora Draper. The tradeoff fallacy: How marketers are misrepresenting american consumers and opening them up to exploitation. Report from the Annenberg School for Communication, June 2015. `https://www.asc.upenn.edu/sites/default/files/TradeoffFallacy_1.pdf`.

[86] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In *Proc. of the Workshop on Usable Security*, 2012.

[87] Narseo Vallina-Rodriguez, Jay Shah, Alessandro Finamore, Yan Grunenberger, Konstantina Papagiannaki, Hamed Haddadi, and Jon Crowcroft. Breaking for commercials: Characterizing mobile advertising. In *Proc. of IMC*, 2012.

[88] Robert J. Walls, Eric D. Kilmer, Nathaniel Lageman, and Patrick D. McDaniel. Measuring the impact and perception of acceptable advertisements. In *Proc. of IMC*, 2015.

[89] Craig E. Wills and Can Tatar. Understanding what they do with what they know. In *Proc. of WPES*, 2012.

[90] Stewart Wolpin. International privacy day: Protect your digital footprint. The Huffington Post, January 2015. `http://www.huffingtonpost.com/stewart-wolpin/international-privacy-day_b_6551012.html`.

[91] Apostolis Zarras, Alexandros Kapravelos, Gianluca Stringhini, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. The dark alleys of madison avenue: Understanding malicious advertisements. In *Proc. of IMC*, 2014.

[92] Sebastian Zimmeck, Jie S. Li, Hyungtae Kim, Steven M. Bellovin, and Tony Jebara. A privacy analysis of cross-device tracking. In *Proc. of USENIX Security Symposium*, Vancouver, BC, 2017.